(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 March 2003 (27.03.2003)

PCT

(10) International Publication Number
WO 03/025745 A2

(51) International Patent Classification[7]: G06F 9/46

(21) International Application Number: PCT/GB02/03690

(22) International Filing Date: 9 August 2002 (09.08.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/953,761 17 September 2001 (17.09.2001) US

(71) Applicant: INTERNATIONAL BUSINESS MACHINES CORPORATION [US/US]; New Orchard Road, Armonk, NY 10504 (US).

(71) Applicant *(for MG only)*: IBM UNITED KINGDOM LIMITED [GB/GB]; PO Box 41, North Harbour, Portsmouth, Hampshire PO6 3AU (GB).

(72) Inventors: BEGUN, Ralph, Murray; 9904 Darnell Court, Raleigh, NC 27615 (US). HUNTER, Steven, Wade; 5709 Dutch Creek Drive, Raleigh, NC 27606 (US). NEWELL, Darryl; 10930 Forbes Creek Drive, Apt. S-104, Kirkland, WA 98033 (US).

(74) Agent: MOSS, Robert, Douglas; IBM United Kingdom Limited, Intellectual Property Law, Hursley Park, Winchester, Hampshire SO21 2JN (GB).

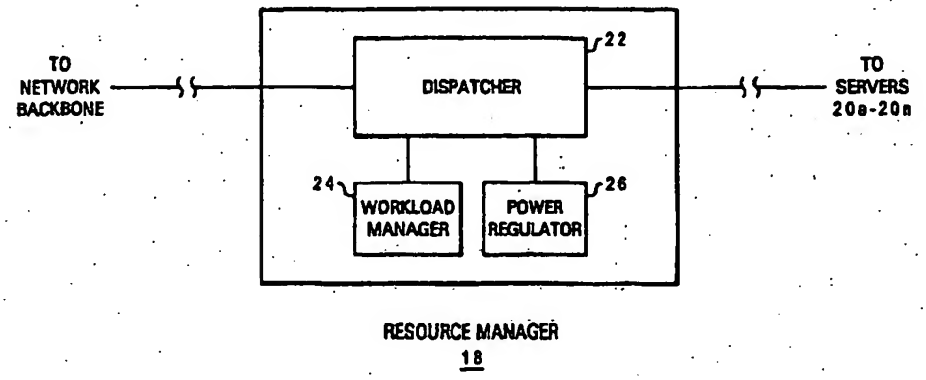
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SYSTEM AND METHOD FOR PERFORMING POWER MANAGEMENT ON A DISTRIBUTED SYSTEM



(57) Abstract: An improved system and method for performing power management on a distributed system. The system utilized to implement the present invention includes multiple servers for processing a set of tasks. The method of performing power management on a system first determines if the processing capacity of the system exceeds a predetermined workload. If the processing capacity exceeds a predetermined level, at least one of the multiple servers on the network is selected to be powered down and the tasks across the remaining servers are rebalanced. If the workload exceeds a predetermined processing capacity of the system and at least a server in a reduced power state may be powered up to a higher power state to increase the overall processing capacity of the system.

WO 03/025745 A2

SYSTEM AND METHOD FOR PERFORMING POWER MANAGEMENT ON A DISTRIBUTED SYSTEM

BACKGROUND OF THE INVENTION

Technical Field

The present invention relates in general to the field of data processing systems, and more particularly, the field of power management in data processing systems. Still more particularly, the present invention relates to a system and method of performing power management on networked data processing systems.

Description of the Related Art

A network (e.g., Internet or Local Area Network (LAN)) in which client requests are dynamically distributed among multiple interconnected computing elements is referred to as a "load sharing data processing system." Server tasks are dynamically distributed in a load sharing system by a load balancing dispatcher, which may be implemented in software or in hardware. Clients may obtain service for requests by sending the requests to the dispatcher, which then distributes the requests to various servers that make up the distributed data processing system.

Initially, for cost-effectiveness, a distributed system may comprise a small number of computing elements. As the number of users on the network increases over time and requires services from the system, the distributed system can be scaled by adding additional computing elements to increase the processing capacity of the system. However, each of these components added to the system also increases the overall power consumption of the aggregate system.

Even though the overall power consumption of a system remains fairly constant for a given number of computing elements, the workload on the network tends to vary widely. The present invention, therefore recognizes that it would be desirable to provide a system and method of scaling the power consumption of the system to the current workload on the network.

SUMMARY OF THE INVENTION

The present invention presents an improved system and method for performing power management for a distributed system. The distributed system utilized to implement the present invention includes multiple servers for processing tasks and a resource manager to determine the relation between the workload and the processing capacity of the system. In response to determining the relation, the resource manager determines whether or not to modify the relation between the workload and the processing capacity of the distributed system.

Accordingly, according to a first aspect the present invention provides a method of performing power management on a distributed system The method first determines if the processing capacity of the system exceeds a predetermined workload. If the processing capacity exceeds the workload, at least one of the multiple servers of the system is selected to be powered down to a reduced power state. Then, tasks are redistributed across the plurality of servers. Finally, the selected server(s) is powered down to a reduced power state.

Preferably the method also determines if the workload exceeds a predetermined processing capacity of the system. If so, at least one server in a reduced power state may be powered up to a higher power state to increase the overall processing capacity of the system. Preferably tasks are then redistributed across the servers in the system.

According to a second aspect the present invention provides a resource manager for performing power management in a distributed system comprising a plurality of servers, the resource manager comprising: a means for receiving a plurality of tasks and relaying said tasks to said distributed system; means for balancing said tasks on said distributed system; means for determining whether or not processing capacity of said distributed system exceeds a current workload; and means, responsive to determining said processing capacity of said distributed system exceeds said current workload, said for selecting and powering down at least one of said plurality of servers to a reduced power state.

Preferably the resource manager further comprises means for determining whether or not said current workload exceeds said processing capacity of said distributed system; and means, responsive to determining said current workload exceeds said processing capacity of said system, for powering up at least one of said plurality of servers to a higher power

state. In this case, preferably the resource manger further comprises means for redistributing said tasks across said plurality of servers.

For example, a dispatcher can provide the means for receiving a plurality of tasks and relaying the tasks to the distributed system, a workload manager (WLM) can provide the means for balancing tasks, and a power regulator can provide the means for determining whether or not processing capacity of a system exceeds a current workload and the means, responsive to this, for selecting and powering down at least one of the plurality of servers to a reduced power state.

Alternatively, for example, an interactive session support (ISS) can provide the means for determining whether or not processing capacity of a system exceeds a current workload, a power manager can provide the means, responsive to the ISS determining this, for selecting and powering down at least one of the plurality of servers to a reduced power state, and a dispatcher can provide the means for balancing tasks , under the control of switching logic. The current workload can be associated with, for example, a plurality of tasks. In this example, preferably the ISS further provides means for determining whether or not the current workload exceeds the processing capacity of the distributed system, and the power regulator further provides, means, responsive to the ISS determining this, for powering up at least one of the plurality of servers to a higher power state.

According to a third aspect the invention provides a distributed data processing system comprising the resource manager of the second aspect and a plurality of servers for processing tasks relayed from said resource manager.

According to a fourth aspect the present invention provides a computer program product comprising instructions which, when executed on a data processing host, cause the host to carry out a method according to the first aspect.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention will now be described, by way of example only, with reference to preferred embodiments thereof, as illustrated in the accompanying drawings, in which:

Figure 1 illustrates an exemplary distributed system that may be utilized to implement a first preferred embodiment of the present invention;

Figure 2 depicts a block diagram of a resource manager utilized for load balancing and power management according to a first preferred embodiment of the present invention;

Figure 3 illustrates an exemplary distributed system that may be utilized to implement a second preferred embodiment of the present invention.

Figure 4 depicts a block diagram of a resource manager utilized for load balancing according to a second preferred embodiment of the present invention;

Figure 5 illustrates a connection table utilized for recording existing connections according to a second preferred embodiment of the present invention;

Figure 6 depicts a layer diagram for the software, including a power manager, utilized to implement a second preferred embodiment of the present invention; and

Figure 7 illustrates a high-level logic flowchart depicting a method for performing power management for a system according to both a first and second preferred embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

The following description of the preferred embodiments of the present invention utilizes the following terms:

"Input/output (I/O) utilization" can be determined by monitoring a pair of queues (or buffers) associated with one or more I/O port(s). A first queue is the receive (input) queue, which temporarily stores data awaiting processing. A second queue is the transmit (output) queue, which temporarily stores data awaiting transmission to another location. I/O utilization can also be determined by monitoring transmit control protocol (TCP) flow and/or congestion control, which indicates the conditions of the network, and/or system.

"Workload" is defined as the amount of (1) I/O utilization, (2) processor utilization, or (3) any other performance metric of servers employed to process or transmit a data set.

"Throughput" is the amount of workload performed in a certain amount of time.

"Processing capacity" is the configuration-dependent maximum level of throughput.

"Reduced power state" is the designated state of a server operating at a relatively lower power mode. There may be several different reduced power states. A data processing system can be completely powered off and require a full reboot of the hardware and operating system. The main disadvantage of this state is the latency required to perform a full reboot of the system. A higher power state is a "sleep state," in which at least some data processing system components (e.g., direct access storage device (DASD), memory, and buses) are powered down, but can be brought to full power without rebooting. Finally, the data processing system may be in a higher power "idle state," with a frequency throttled processor, inactive DASD, but the memory remains active. This state allows the most rapid return to a full power state and is therefore employed when a server is likely to be idle for a short duration.

"Reduced power server(s)" is a server or group of servers operating in a "reduced power state."

"Higher power state" is the designated state of a server operating at a relatively higher power than a reduced power state.

"Higher power server(s)" is a server or group of servers operating in a "higher power state."

"Frequency throttling" is a technique for changing power consumption of a system by reducing or increasing the operational frequency of a system. For example, by reducing the operating frequency of the processor under light workload requirements, the processor (and system) employs a significantly less amount of power for operation, since power consumed is related to the power supply voltage and the operating frequency.

In one embodiment of the present invention, data processing systems communicate by sending and receiving Internet protocol (IP) data requests

via a network such as the Internet. IP defines data transmission utilizing data packets (or "fragments"), which include an identification header and the actual data. At a destination data processing system, the fragments are combined to form a single data request.

With reference now to the figures, and in particular, with reference to Figure 1, there is depicted a block diagram of a network 10 in which a first preferred embodiment of the present invention may be implemented. Network 10 may be a local area network (LAN) or a wide area network (WAN) coupling geographically separate devices. Multiple terminals 12a-12n, which can be implemented as personal computers, enable multiple users to access and process data. Users send data requests to access and/or process remotely stored data through network backbone 16 (e.g., Internet) via a client 14.

Resource manager 18 receives the data requests (in the form of data packets) via the Internet and relays the requests to multiple servers 20a-20n. Utilizing components described below in more detail, resource manager 18 distributes the data requests among servers 20a-20n to promote (1) efficient utilization of server processing capacity and (2) power management by powering down selected servers to a reduced power state when the processing capacity of servers 20a-20n exceeds a current workload.

During operation, the reduced power state selected depends greatly on the environment of the distributed system. For example, in a power scarce environment, uneeded servers can be completely powered off. Such an implementation may be appropriate for a power sensitive distributed system where response time is not critical.

Also, if the response time is critical to the operation of the distributed system, a full shutdown of unneeded servers and the subsequent required reboot time might be undesirable. In this case, the selected reduced power state might only be the frequency throttling of the selected unneeded server or even the "idle state." In both cases, the reduced power servers may be quickly powered up to meet the processing demands of the data requests distributed by resource manager 18.

Referring to Figure 2, there is illustrated a detailed block diagram of resource manager 18 according to a first preferred embodiment of the present invention. Resource manager 18 may comprise a dispatcher component 22 for receiving and sending data requests to and from servers

20a-20n to prevent any single higher power server's workload from exceeding the server's processing capacity.

Preferably, a workload management (WLM) component 24 determines a server's processing capacity utilizing more than one performance metric, such as I/O utilization and processor utilization, before distributing data packets over servers 20a-20n. In certain transmission-heavy processes, five percent of the processor may be utilized, but over ninety percent of the I/O may be occupied. If WLM 24 utilized processor utilization as its sole measure of processing capacity, the transmission-heavy server may be wrongfully powered down to a reduced power state when powering up a reduced power server to rebalance the transmission load might be more appropriate. Therefore, WLM 24 or any other load balancing technology used to implement an embodiment of the present invention preferably monitors at least (1) processor utilization, (2) I/O utilization, and (3) any other performance metric (also called a "custom metric"), which may be specified by a user.

After determining the processing capacity of servers 20a-20n, WLM 24 selects a server best suited for receiving a data packet. Dispatcher 22 distributes the incoming data packets to the selected server by (1) examining identification field of each data packet, (2) replacing the address in destination address field with an address unique to the selected server, and (3) relaying the data packet to the selected server.

Power regulator 26 operates in concert with WLM 24 by monitoring incoming and outgoing data to and from servers 20a-20n. If a higher power server remains idle (e.g., does not receive or send a data request for a predetermined interval) or available processing capacity exceeds a workload, determined by a combination of I/O utilization, processor utilization, and any other custom metric, WLM 24 selects at least one higher power server to power down to a reduced power state. If the selected reduced power state is a full power down or sleep modes, dispatcher 22 redistributes the tasks (e.g., functions to be performed by the selected higher power server) on the higher power servers selected for powering down among the remaining higher power servers and sends a signal that indicates to power regulator 26 that dispatcher 22 has completed the task redistribution. Then, power regulator 26 powers down a higher power server to a reduced power state.

If the selected reduced power state is an idle or frequency throttled state, dispatcher 22 redistributes a majority of the tasks on

the higher power severs selected for powering down among the higher power servers. However, the frequency throttled server may still process tasks, but at a reduced capacity. Therefore, some tasks remain on the frequency throttled server despite its reduced power state.

If the tasks on the higher power servers exceeds the processing capacity, power regulator 26 powers up a reduced power server, if available, to a higher power state to increase the processing capacity of servers 20a-20n. Dispatcher 22 redistributes the tasks across the new set of higher power servers to take advantage of the increase processing capacity.

An advantage to this first preferred embodiment of the present invention is the more efficient power consumption of the distributed server. If the processing capacity of the system exceeds the current workload, at least one higher power server may be powered down to a reduced power state, thus decreasing the overall power consumption of the system.

One drawback to this first preferred embodiment of the present invention is the installation of resource manager 18 as a bidirectional passthrough device between the network and servers 20a-20n, which may result in a significant bottleneck in networking throughput from the servers to the network. The user of a single resource manager 18 also creates a single point of failure between the server group and the client.

With reference to Figure 3, there is depicted a block diagram of a network 30 in which a second preferred embodiment of the present invention may be implemented. Network 30 may also be a local area network (LAN) or a wide area network (WAN) coupling geographically separate devices. Multiple terminals 12a-12n, which can be implemented as personal computers, enable multiple users to access and process data. Users send data requests for remotely stored data through a client 14 and a network backbone 16, which may include the Internet. Resource manager 28 receives the data requests via the Internet and relays the data request to a dispatcher (32 of Figure 4), which assigns each data request to a specific server. Unlike the first preferred embodiment of the present invention, servers 20a-20n sends outgoing data packets directly to client 14 via network backbone 16, instead of sending the data packet back through dispatcher 32.

Referring to Figure 4, there is illustrated a block diagram of resource manager 28 according to a second preferred embodiment of the present invention. Dispatcher 32, coupled to a switching logic 34, distributes tasks received from network backbone 16 to servers 20a-20n. Dispatcher 32 examines each data request identifier in each data packet identification header and compares the identifier to other identifiers listed in an identification field 152 in a connection table (as depicted in Figure 5) stored in memory 36. Referring to Figure 5, connection table 150 includes two fields: identification field 152 and a corresponding assigned server field 154. Identification field 152 lists existing connections (e.g., pending data requests) and assigned server field 154 indicates the server assigned to the existing connection. If the data request identifier from a received data packet matches another identifier listed on connection table 150, the received data packet represents an existing connection, and dispatcher 32 automatically forwards to the appropriate server the received data packet utilizing the server address in an assigned server field 154. However, if the data request identifier does not match another identifier listed on connection table 150, the data packet represents a new connection. Dispatcher 32 records the request identifier from the data packet into identification field 152, selects an appropriate server to receive the new connection (to be explained below in more detail), and records the address of the appropriate server in assigned server field 154.

With reference to Figure 6, there is illustrated a diagram outlining an exemplary software configuration stored in servers 20a-20n according to a second preferred embodiment of the present invention. As well-known in the art, a data processing system (e.g., servers 20a-20n) requires an operating system, to function properly. Basic functions (e.g., saving data to a memory device or controlling the input and output of data by the user) are handled by operating system 50, which may be at least partially stored in memory and/or direct access storage device (DASD) of the data processing system. A set of application programs 60 for user functions (e.g., an e-mail program, word processors, Internet browsers) runs on top of operating system 50. As shown, interactive session support (ISS) 54, and power manager 56 access the functionality of operating system 50 via an application program interface (API) 52.

ISS (Interactive Session Support) 54, a domain name system (DNS) based component installed on each of servers 20a-20n, implements I/O utilization, processor utilization, or any other performance metric (also

called a "custom metric") to monitor the distribution of the tasks over servers 20a-20n. Functioning as an "observer" interface that enables other applications to monitor the load distribution, ISS 54 enables program manager 56 to power up or power down servers 20a-20n as workload and processing capacities fluctuate. Dispatcher 32 also utilizes performance metric data from ISS 54 to perform load balancing functions for the system. In response to receiving a data packet representing a new connection, dispatcher 32 selects an appropriate server to assign a new connection utilizing task distribution data from ISS 54.

Power manager 56 operates in concert with dispatcher (32 of Figure 4) via ISS 54 by monitoring incoming and outgoing data to and from servers 20a-20n. If a higher power server remains idle (e.g., does not receive or send a data request for a predetermined time) or available processing capacity exceeds a predetermined workload, as determined by ISS 54, dispatcher 32 selects a higher power server to be powered down to a reduced power state, redistributes the tasks of among the remaining higher power servers and sends a signal to power manager 56 indicating the completion of task redistribution. Power manager 56 powers down the selected higher power server to a reduced power state, in response from receiving the signal from dispatcher 32. Also, if the workload on the higher power servers exceeds the processing capacity, power manager 56 powers up a reduced power server, if available, to a higher power state to increase the processing capacity of servers 20a-20n. Dispatcher 32 then redistributes the tasks among the new set of higher power servers to take advantage of the increased processing capacity.

Referring now to Figure 7, there is depicted a high-level logic flowchart depicting a method of power management. A first preferred embodiment of the present invention can implement the method utilizing resource manager 18, which includes power regulator 26, for controlling power usage in servers 20a-20n, workload manager (WLM) 24, and dispatcher 22 for dynamically distributing the tasks over servers 20a- 20n. A second preferred embodiment of the present invention utilizes a resource manager that includes dispatcher 32, ISS 54, and power manager 56 to manage power usage in servers 20a-20n. These components can be implemented in hardware, software and/or firmware as will be appreciated by those skilled in the art.

In the following method, all rebalancing functions are performed by WLM 24 and dispatcher 22 in the first preferred embodiment (Figure 2) and

dispatcher 32 in the second preferred embodiment (Figure 4). All determinations, selection, and powering functions employ power regulator 26 in the first preferred embodiment and power manager 56 and ISS 54 in the second preferred embodiment.

As illustrated in Figure 7, the process begins at block 200, and enters a workload analysis loop, including blocks 204, 206, 208, and 210. At block 204, a determination is made of whether or not the aggregate processing capacity of servers 20a-20n exceeds a current workload. The current workload is determined utilizing server performance metrics (e.g., processor utilization and I/O utilization) and compared to the current processing capacity of servers 20a-20n.

If the processing capacity of servers 20a-20n exceeds the current workload, the process continues to block 206, which depicts the selection of at least a server to be powered down to a reduced power state. The total tasks on servers 20a-20n are rebalanced across the remaining servers, as depicted at block 208. As illustrated in block 210, the selected server(s) is powered down to a reduced power state. Finally, the process returns from block 210 to block 204.

As depicted at block 212, a determination is made of whether or not the workload exceeds the processing capacity of servers 20a-20n. If the workload exceeds the processing capacity of servers 20a-20n, at least a server is selected to be powered up to a higher power state, as illustrated in block 214. At least the selected server(s) is powered up, as depicted in block 216, and the tasks is rebalanced over servers 20a-20n. The process returns from block 218 to block 204, as illustrated.

The preferred embodiments of the present invention implement a resource manager coupled to a group of servers. The resource manager analyzes the balance of tasks of the group of servers utilizing a set of performance metrics. If the processing capacity of the group of higher power servers exceeds current workload, at least a server in the group is selected to be powered down to a reduced power state. The tasks on the selected server are rebalanced over the remaining higher power servers. However, if the power manager determines that the workload exceeds the processing capacity of the group of servers, at least a server is powered up to a higher power state, and the tasks are rebalanced over the group of servers.

CLAIMS

1. A method for power management in a distributed system comprising a plurality of servers, said method comprising:

determining whether or not processing capacity of said system exceeds a current workload associated with a plurality of tasks;

in response to determining said processing capacity of said system exceeds said workload, selecting at least one of said plurality of servers to be powered down to a reduced power state;

rebalancing said tasks across said plurality of servers; and

powering down said at least one selected server to a reduced power state.

2. The method according to claim 1, further including:

determining whether or not said workload exceeds said processing capacity of said system; and

in response to determining said workload exceeds said processing capacity of said system, powering up at least one of said plurality of servers to a higher power state.

3. The method according to claim 2, further comprising:

redistributing said tasks across said plurality of servers.

4. A resource manager for performing power management in a distributed system, the distributed system comprising a plurality of servers, the resource manager comprising:

a means for receiving a plurality of tasks and relaying said tasks to said distributed system;

a means for balancing said tasks on said distributed system;

a means for determining whether or not processing capacity of said distributed system exceeds a current workload; and

means responsive to determining said processing capacity of said distributed system exceeds said current workload, for selecting and powering down at least one of said plurality of servers to a reduced power state.

5. A resource manager of claim 4, further comprising:

means for determining whether or not said current workload exceeds said processing capacity of said distributed system; and

means, responsive to determining said current workload exceeds said processing capacity of said system, for powering up at least one of said plurality of servers to a higher power state.

6. A resource manager of claim 5 further comprising:

means for redistributing said plurality of tasks across said plurality of servers.

7. A distributed data processing system, comprising:

a resource manager in accordance with any one of claims 4 to 6; and

a plurality of servers for processing tasks relayed from said resource manager.

8. A computer program product comprising instructions, which, when executed in a data processing system, cause said system to carry out a method according to any one of claims 1 to 3.

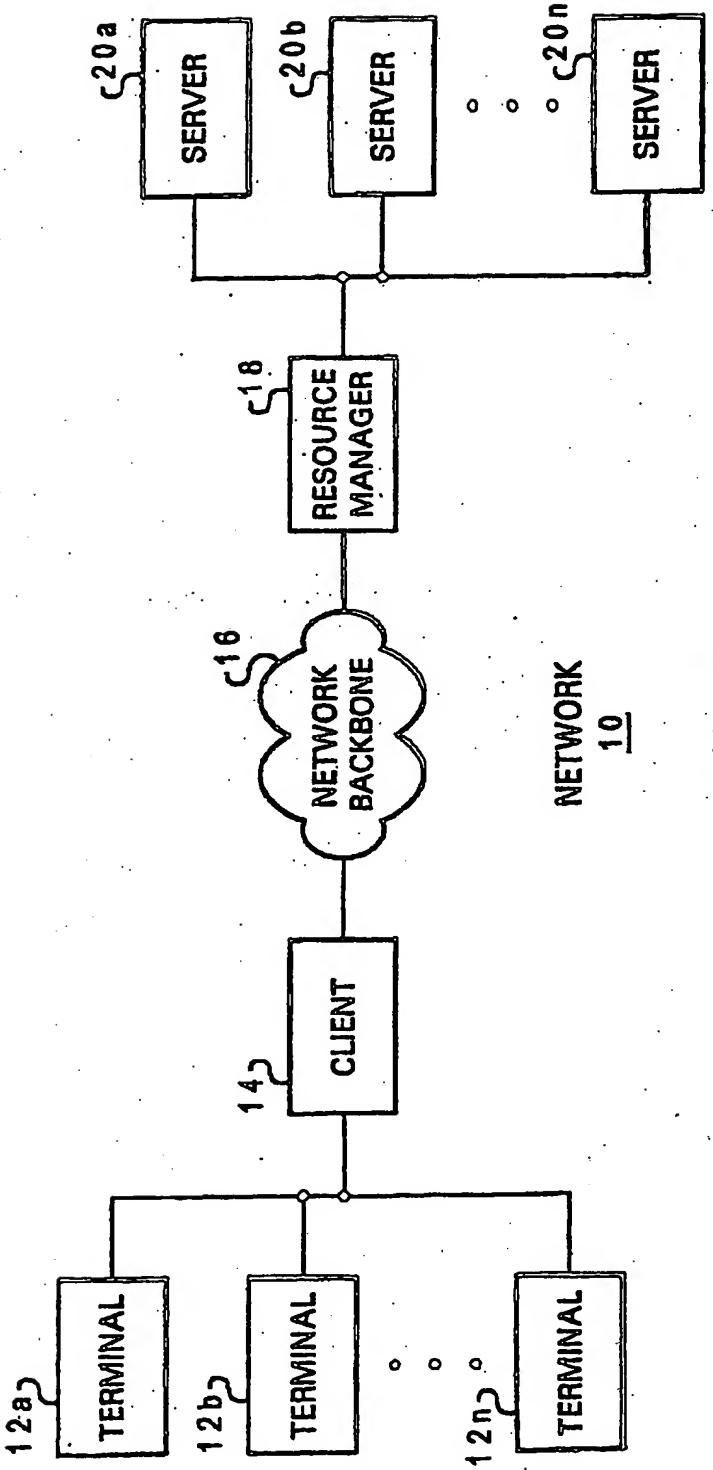

20a
SERVER
20b
SERVER
20n
SERVER
18
RESOURCE MANAGER
16
NETWORK BACKBONE
14
CLIENT
12a
TERMINAL
12b
TERMINAL
12n
TERMINAL
NETWORK
10

Fig. 1

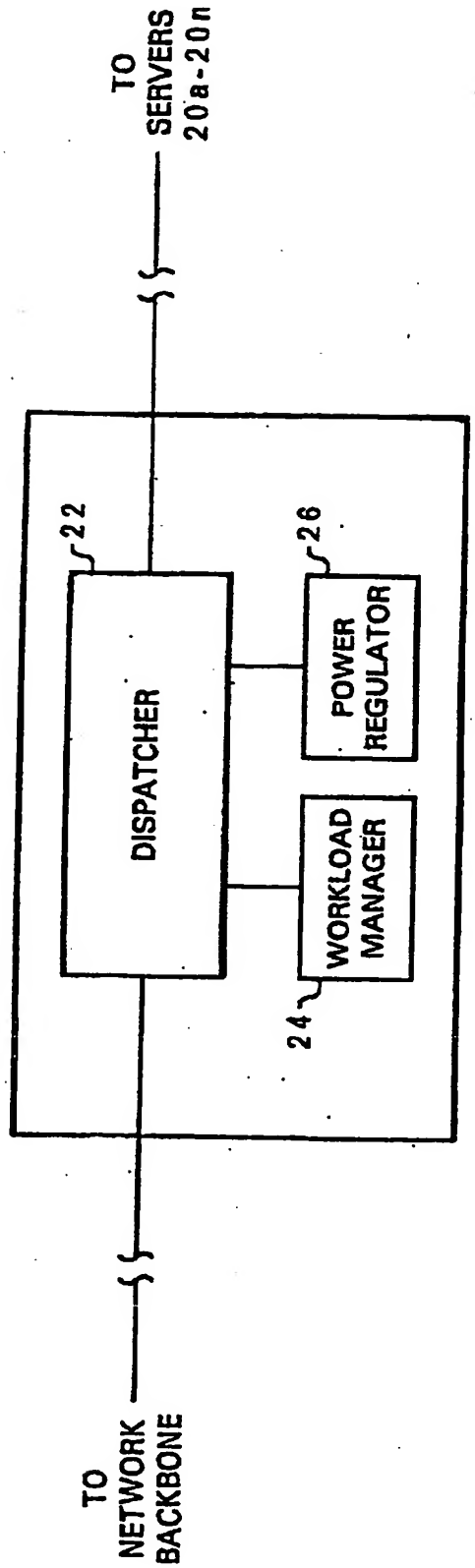

TO SERVERS 20a-20n
22
26
DISPATCHER
POWER REGULATOR
WORKLOAD MANAGER
24
RESOURCE MANAGER 18
TO NETWORK BACKBONE

Fig. 2

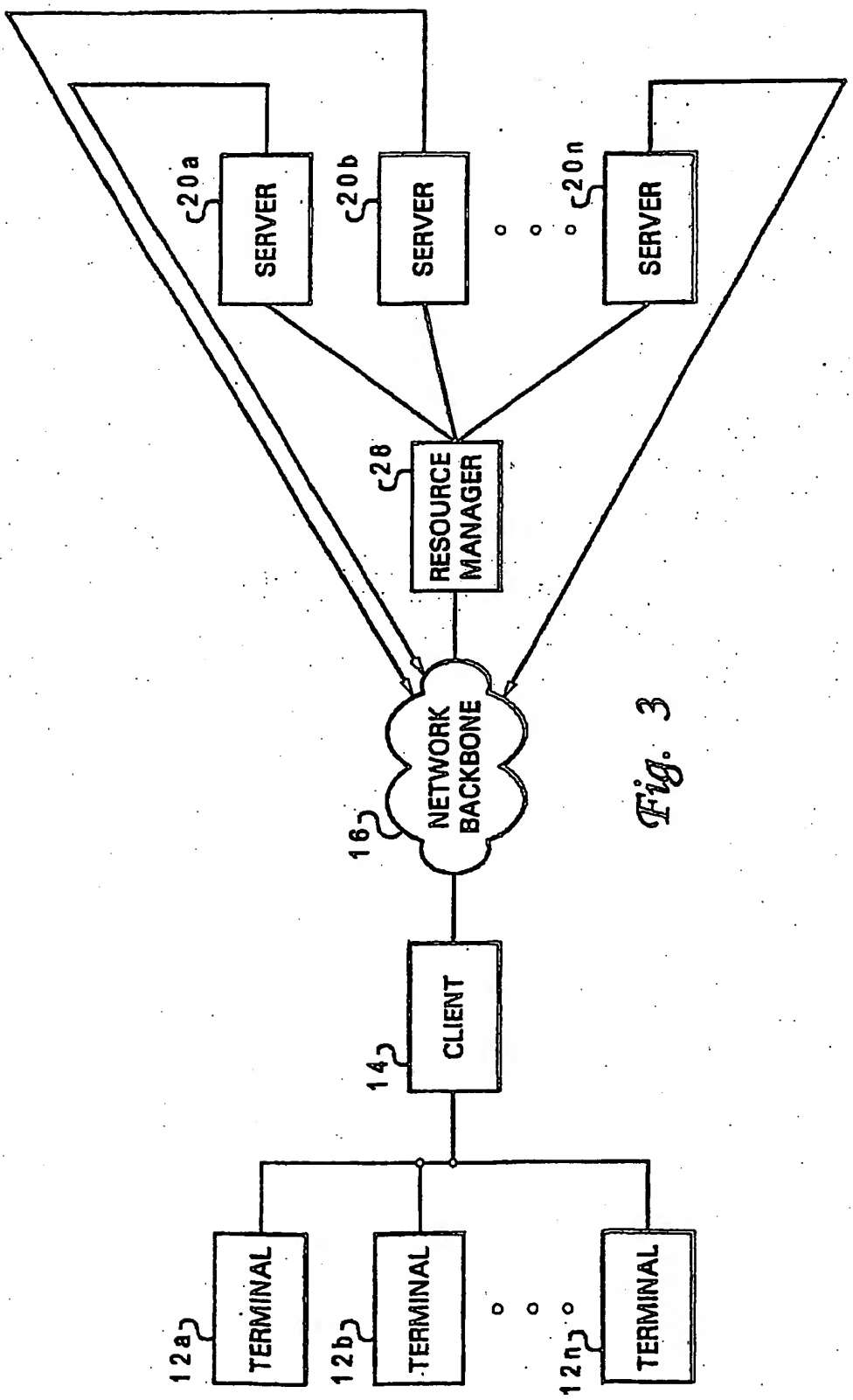

SERVER
20a
SERVER
20b
SERVER
20n
RESOURCE MANAGER
28
NETWORK BACKBONE
16
CLIENT
14
TERMINAL
12a
TERMINAL
12b
TERMINAL
12n

Fig. 3

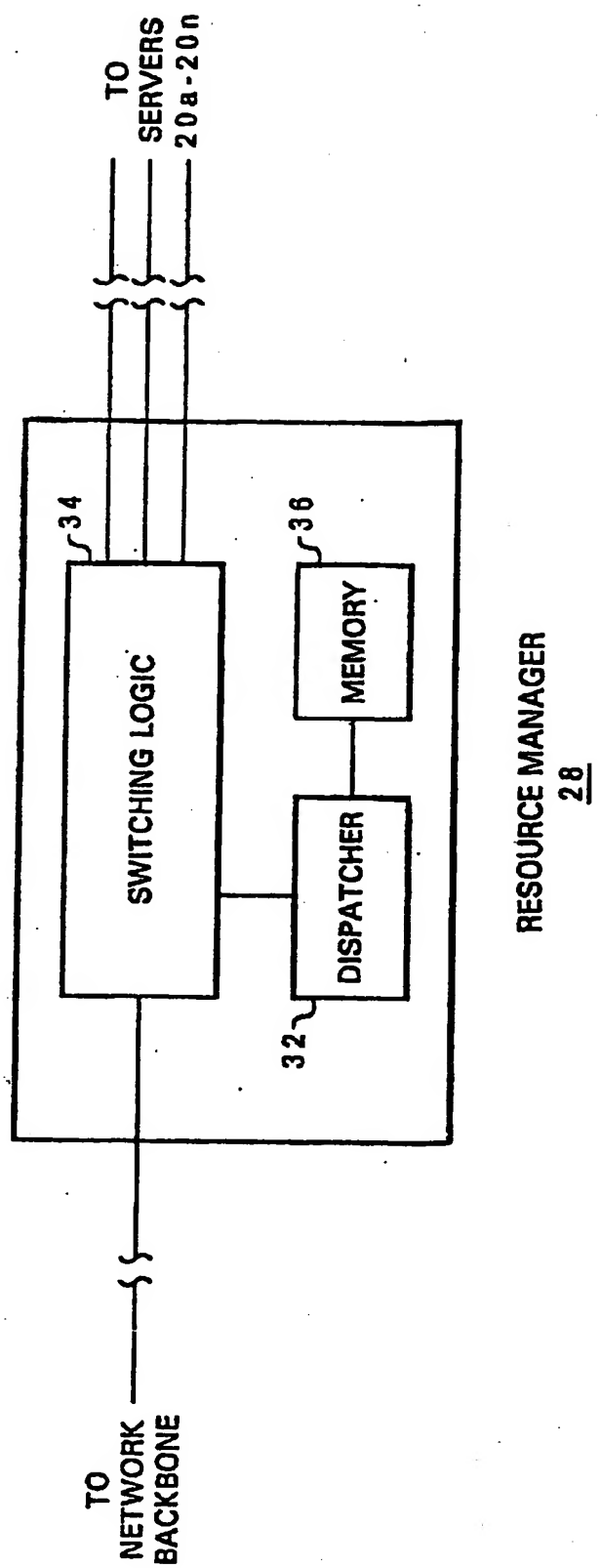

TO SERVERS 20a-20n
34
36
SWITCHING LOGIC
MEMORY
DISPATCHER
32
RESOURCE MANAGER 28
TO NETWORK BACKBONE

Fig. 4

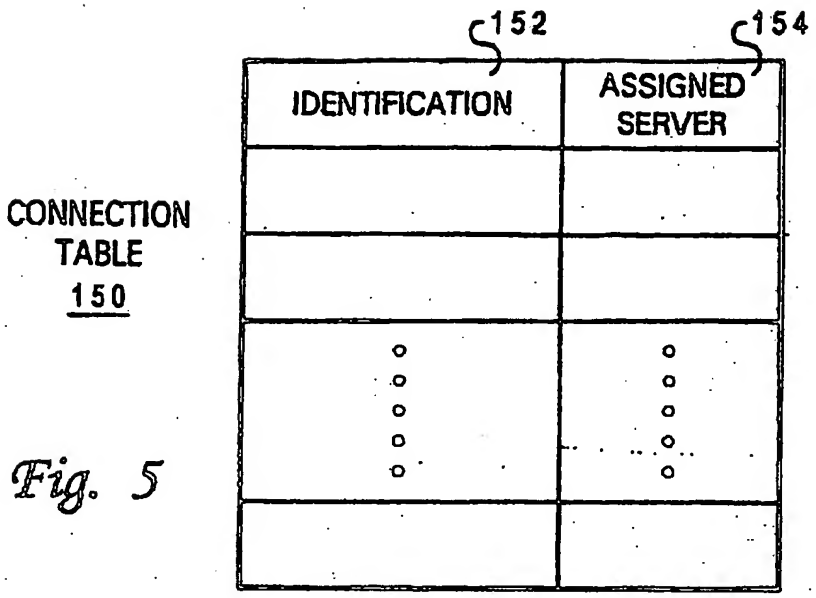

152
154
IDENTIFICATION
ASSIGNED SERVER
CONNECTION TABLE
150

Fig. 5

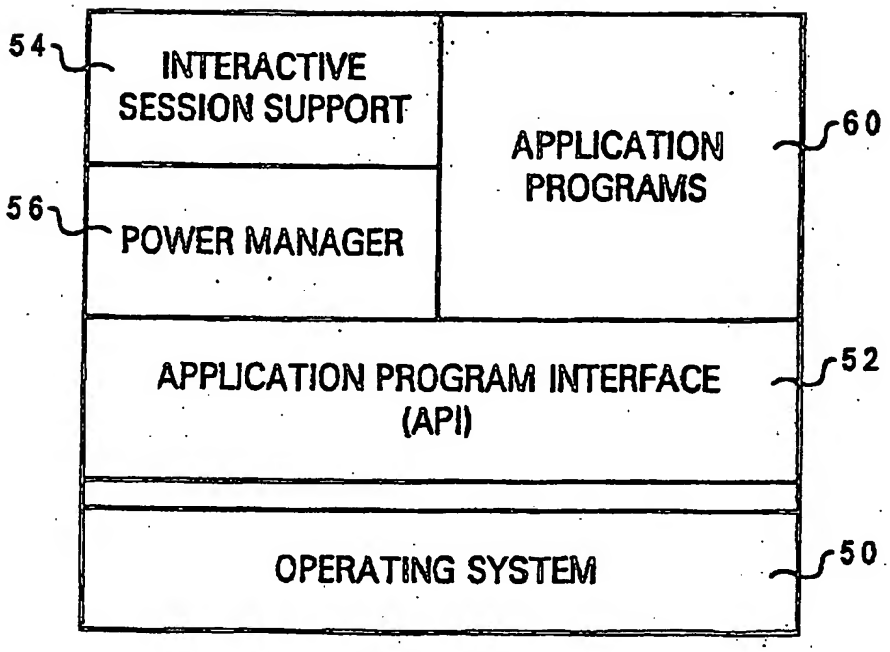

54
INTERACTIVE SESSION SUPPORT
56
POWER MANAGER
APPLICATION PROGRAMS
60
APPLICATION PROGRAM INTERFACE (API)
52
OPERATING SYSTEM
50

Fig. 6

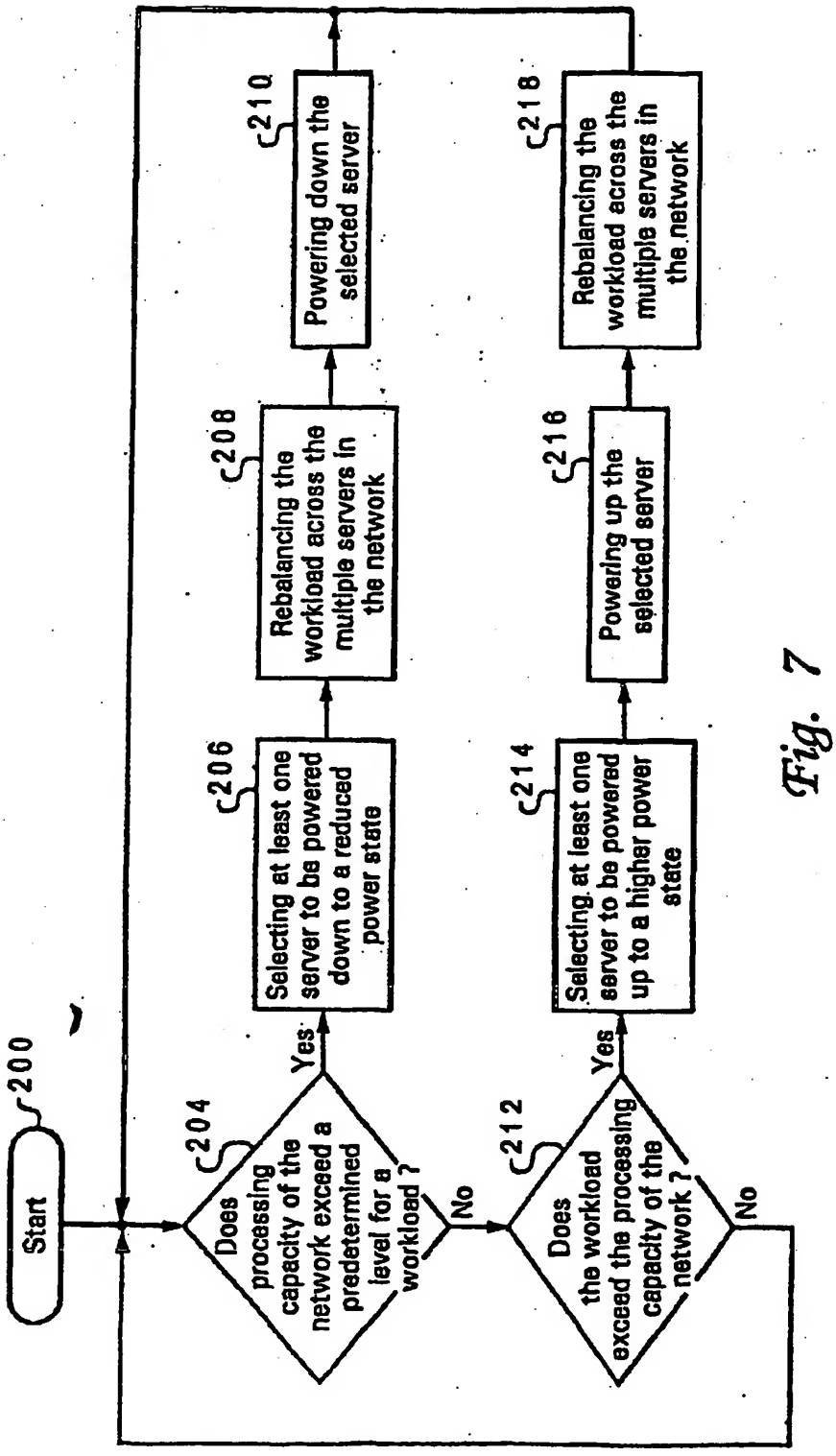

Start
200
Does processing capacity of the network exceed a predetermined level for a workload ?
204
Yes
No
Selecting at least one server to be powered down to a reduced power state
206
Rebalancing the workload across the multiple servers in the network
208
Powering down the selected server
210
Does the workload exceed the processing capacity of the network ?
212
Yes
No
Selecting at least one server to be powered up to a higher power state
214
Powering up the selected server
216
Rebalancing the workload across the multiple servers in the network
218

Fig. 7

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 March 2003 (27.03.2003)

PCT

(10) International Publication Number
WO 2003/025745 A3

(51) International Patent Classification[7]: G06F 9/50, 1/32

(21) International Application Number:
PCT/GB2002/003690

(22) International Filing Date: 9 August 2002 (09.08.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/953,761 17 September 2001 (17.09.2001) US

(71) Applicant: INTERNATIONAL BUSINESS MACHINES CORPORATION [US/US]; New Orchard Road, Armonk, NY 10504 (US).

(71) Applicant *(for MG only)*: IBM UNITED KINGDOM LIMITED [GB/GB]; PO Box 41, North Harbour, Portsmouth, Hampshire PO6 3AU (GB).

(72) Inventors: BEGUN, Ralph, Murray; 9904 Darnell Court, Raleigh, NC 27615 (US). HUNTER, Steven, Wade; 5709 Dutch Creek Drive, Raleigh, NC 27606 (US). NEWELL, Darryl; 10930 Forbes Creek Drive, Apt. S-104, Kirkland, WA 98033 (US).

(74) Agent: MOSS, Robert, Douglas; IBM United Kingdom Limited, Intellectual Property Law, Hursley Park, Winchester, Hampshire SO21 2JN (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

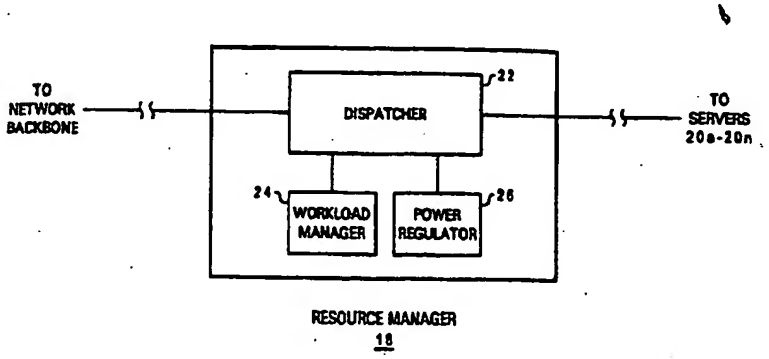
(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report:
19 February 2004

[Continued on next page]

(54) Title: SYSTEM AND METHOD FOR PERFORMING POWER MANAGEMENT ON A DISTRIBUTED SYSTEM



(57) Abstract: An improved system and method for performing power management on a distributed system. The system utilized to implement the present invention includes multiple servers for processing a set of tasks. The method of performing power management on a system first determines if the processing capacity of the system exceeds a predetermined workload. If the processing capacity exceeds a predetermined level, at least one of the multiple servers on the network is selected to be powered down and the tasks across the remaining servers are rebalanced. If the workload exceeds a predetermined processing capacity of the system and at least a server in a reduced power state may be powered up to a higher power state to increase the overall processing capacity of the system.

WO 2003/025745 A3

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

INTERNATIONAL SEARCH REPORT

PCT/GB 02/03690

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F9/50 G06F1/32

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	E. PINHEIRO ET AL: "Load Balancing and Unbalancing for Power and Performance in Cluster-Based Systems" TECHNICAL REPORT DCS-TR-440, 'Online! May 2001 (2001-05), pages 4.1-4.8, XP002261813 Rutgers University, New Jersey, USA Retrieved from the Internet: <URL:http://research.ac.upc.es/pact01/colp /paper04.pdf> 'retrieved on 2003-11-13! page 4.1, right-hand column, line 3 - line 18 page 4.2, left-hand column, line 10 -page 4.3, left-hand column, line 24 page 4.6, left-hand column, line 16 - line 35 —— -/—	1-8

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 November 2003	05/01/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax (+31-70) 340-3016	Michel, T

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

PCT/GB 02/03690

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	J. CHASE AND R. DOYLE: "Balance of Power: Energy Management for Server Clusters" IN PROCEEDINGSOF THE 8TH WORKSHOPON HOT TOPICS IN OPERATING SYSTEMS, 'Online! May 2001 (2001-05), pages 1-6, XP002261814 Retrieved from the Internet: <URL:http://citeseer.nj.nec.com/rd/0%2C420 187%2C1%2C0.25%2CDownload/http://citeseer. nj.nec.com/cache/papers/cs/20292/http:zSzz Szwww.cs.duke.eduzSzarizSzpublicationszSzb alance-of-power.pdf/chase01balance.pdf> 'retrieved on 2003-11-12! the whole document	1-8
A	EP 0 978 781 A (LUCENT TECHNOLOGIES INC) 9 February 2000 (2000-02-09) abstract	1,4
P,X	CHASE J S ET AL: "Managing energy and server resources in hosting centers" 18TH ACM SYMPOSIUM ON OPERATING SYSTEMS PRINCIPLES (SOSP'01), BANFF, ALTA., CANADA, 21-24 OCT. 2001, vol. 35, no. 5, pages 103-116, XP002261815 Operating Systems Review, Dec. 2001, ACM, USA ISSN: 0163-5980 the whole document	1-8
T	D. BRADLEY ET AL: "Workload-based power management for parallel computer systems" IBM JOURNAL RESEARCH AND DEVELOPMENT, vol. 47, no. 5/6, September 2003 (2003-09), pages 703-718, XP002261816 the whole document	

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

PCT/GB 02/03690

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0978781 A	09-02-2000	US 6141762 A	31-10-2000
		EP 0978781 A2	09-02-2000
		JP 2000066776 A	03-03-2000